下载APP

08 | 标准差：这人是不是“靠谱”其实看标准差？

2021-08-13 郭炜

《数据分析思维课》 课程介绍 >

讲述：郭炜

时长 15:22 大小 14.09M

数据给你一双看透本质的眼睛，这里是《数据分析思维课》，我是郭炜。

前面我们讲过平均值不能够代表整体的水平，也给你讲了大数定律、散点图这些知识。接下来我们再进入一个常见的问题：**怎样能快速看清一组数据的大概情况？**

对于这个问题来说，我们不必用非常复杂的散点图或者文字来进行表述，这个时候就轮到标准差登场了。标准差和数据分布、平均值一起就可以很方便地描述一组数据的大致情况。

标准差还有一个孪生兄弟叫做标准误差，这两个兄弟确实很像，我们也会经常能听到说“这个问题在误差允许的范围里。”感觉一旦说了这句话，好像这个东西就很靠谱了，但真的这样吗？今天我就给你展开讲讲标准差和标准误差。

标准差

标准差的概念比较简单，它代表一组数值和平均值相比分散开来的程度。也就是说，标准差大代表大部分的数值和平均值差异比较大，标准差小代表这组数字比较接近平均值。

标准差的计算公式我给你放在了附录里，公式看上去稍微复杂一点，但主要就是算每一个数据和平均值之间的差异距离。你经常听说某市平均薪资是 X 万，你很纳闷我和周边人薪资这么少，为啥平均薪资那么高，我是怎么“被涨薪”的？我们可以看一下下面这个例子，假如两个小组的月薪大概是如下这个样子，单位都是“万”。

第一组：[1.72，1.70，1.68，1.71，1.69] ；第二组：[1.70，5.20，0.60，0.2，0.8]。

这两组人你可以简单计算一下，你会发现平均月薪都是 1 万 7。但很明显，第 2 组人的薪酬高低差异要比第 1 组人大很多。第一组人都是 1 万 7 左右的薪资，差异不大。你很不巧在第二组里，你月薪 6000，周边都是 2000、8000 的小伙伴，但是实际上，你这组里有月薪 5 万的人你不认识，于是，你就“被涨薪”了。

通过公式或者 Excel 函数（我在最后一章会教你如何方便地计算），你能算出来第一组标准差是 0.014，第二组是 1.818，差异能有一百多倍。如果每次只给你某地区或者某部门的平均薪酬，但是不告诉你这个地区部门它的标准差有多大，那我们难免就会觉得困惑，“不患寡而患不均”用在这里依旧很合适。

所以看薪资的时候，**你不仅需要知道一个平均值，同时还需要知道一个标准差**，你才能知道整体薪资水平、你自己的水平以及你将来的天花板在哪里。

但是只有这个概念还不够，假设对于第 1 组的薪资单位来说，我用的是不是“万元”，而是用“百元”甚至“元”作为单位的话，它的标准差就会到 1.414 和 141.4。这个时候再和第 2 组人员去比，感觉好像标准差的离散度更高，但是实际数据却不是如此。

所以一般我们真的在做数据分析的时候，我们会常用另外一个数据来规避这种问题，它叫做离散系数 CV（coefficient of variation）。它的计算公式很简单，就是用标准差除以平均值（**离散系数 = 标准差 / 平均值**），这样的话就规避了单位或者其他因素的这些差异。我们直接看离散系数这个数据，就能知道这几组数据之间的离散程度和差异是什么样的。

下次你再去问人力资源部门的平均薪酬的时候，你可以多问一句“这个部门的离散系数有多少？”你大概就会知道，你可以要到最高多少的薪酬和你将来的涨薪空间会有多大了。

标准差的具体使用

标准差除了衡量一个群体里面具体数值之间差异有多大，比如说衡量我们的薪酬、身高、体重这些差异之外，它还有什么用呢？

它也会用于衡量一个人或者一个团队的稳定性，比如说。在你常见的 NBA 里我们会用平均数据来衡量一个球员的战斗力，比如场均得分，盖帽，抢断助攻等等。

排名	赛季	球员	球队	上场次数	两分球	三分球	罚球	总得分	平均得分
1	1961-62	威尔特·张伯伦	费城勇士	80	1,597	—	835	4,029	**50.4**
2	1962-63	威尔特·张伯伦	旧金山勇士	80	1,463	—	660	3,586	**44.8**
3	1960-61	威尔特·张伯伦	费城勇士	79	1,251	—	531	3,033	**38.4**
4	1961-62	埃尔金·贝勒	洛杉矶湖人	48	680	—	476	1,836	**38.3**
5	1959-60	威尔特·张伯伦	费城勇士	72	1,065	—	577	2,707	**37.6**
6	1986-87	迈克尔·乔丹	芝加哥公牛	82	1,098	12	833	3,041	**37.1**
7	1963-64	威尔特·张伯伦	旧金山勇士	80	1,024	—	540	2,948	**36.9**
8	2018-19	詹姆斯·哈登	休斯敦火箭	76	822	368	737	2,749	**36.2**
9	1966-67	里克·巴里	旧金山勇士	78	1,011	—	753	2,775	**35.6**
10	2005-06	科比·布莱恩特	洛杉矶湖人	80	978	180	696	2,832	**35.4**
11	1987-88	迈克尔·乔丹	芝加哥公牛	82	1,069	7	723	2,868	**35.0**
12	1971-72	卡里姆·阿卜杜勒-贾巴尔	密尔沃基雄鹿	81	1,159	—	504	2,822	**34.8**

同时，我们会使用标准差来衡量一个球员的稳定性。

> 如果只看场均得分达到 20+ 的球员中，当属勒布朗 - 詹姆斯最稳定，标准差为 5.8 分，遍历他本赛季所有的比赛，他既没有超过 40 分的狂飙，也无低于 13 分的低迷。

类似的，我们在衡量一个团队的销售业绩整体情况的时候，我们会使用平均值。但是如果我们要看的是一段时间内团队成员的收入稳定情况和能力，我们就会看他最近成单的标准差。

同样，对应到做管理上，比如我做 CTO 管理程序员的时候，我会去留意大家提交代码的节奏。有的人就是喜欢所有事情都到最后一天才完成提交，有的人就喜欢平均用力，在各个时间段里都有提交。

用标准差来看，你就会发现有的人**标准差非常大，属于突击型选手**；有的人**标准差很小，属于细水长流型选手**。对于标准差比较大的人，他的风险就比较高，因为他有可能最后关头完美完成任务，也有可能拖延症拖到最后事情没有完成，最后整体的平均值都没有达到；而按部就班的人，他的标准差比较小，优势是比较稳定，但是突破性可能不够强。学到这里，你也可以试着评估一下你的工作节奏，你是属于哪一种选手呢？

而在做投资的时候，标准差也是一个重要的风险 / 收益衡量指标。你看我们在银行储蓄，这个利率波动就很小，相应地标准差很小；股票的波动就会大一些，收益的标准差也会比较大；你再看比特币，一会儿马斯克一句话翻好几倍，一会儿跌 30%，炒币收益的标准差可能是银行收益标准差的上万倍，股票的上百倍。

所以如果你把钱放在银行，标准差较小，收益稳定；而如果你要去炒币，标准差这么大，你有可能大赚一大笔也可能赔得血本无归。标准差，其实就是代表着一个行业里面的波动情况，特别面对一个你不理解的投资产品的时候，可以看一下这个产品历史的标准差，和你常用的投资品比对一下，你心里就有数了。像黄金这种很稳定的产品，几个标准差就很大了。比如 2013 年 04 月 16 日黄金大跌，路透社分析师约翰·肯普（John Kemp）感叹黄金波动率超过 6 个标准差，觉得非常不可思议。

像黄金一类很少波动的东西出现了这么大的波动，达到 6 个标准差的波动的时候（本来稳定的标准差发生了巨大改变），我们就把这种事件叫做“**黑天鹅事件**”。这次事件也被后来的人叫做“黄金黑天鹅事件”，所以你下次再看到黑天鹅事件的时候，你要知道这个说法是从标准差这里衍生出来的概念。

标准误差

说完标准差，我们来说一下它的孪生兄弟：标准误差。误差这个词我们经常在生活和工作当中提到，说“这个是在我们误差范围里可以接受的。”那么这句话里面提到的“误差范围”说的到底是什么呢？它和标准差是啥关系？

这两个概念经常在很多地方被混淆，以至于很多的统计模型里说的标准差，其实说的是标准误差。这两个概念之间最大的差别其实就是在于，标准差是针对确切一次的已知统计结果，反映的是在一次统计中，个体之间的离散程度，也可以说**标准差是针对具体实例的描述性统计**。

而**标准误差代表一种推论的估计**，它反映的是多次抽样当中样本均值之间的离散程度，也就是反映这次抽样样本均值对于总体期望均值的代表性，它主要是用于推断整体情况预测和推算使用。如果这么说你还是有些分不清这两兄弟，你可以用下面这个两个公式来对照分辨一下。

标准差（Standard deviation）= 一次统计中个体分数间的离散程度，反映了个体对样本整体均值的代表性，用于描述统计。

标准误差（Standard error）= 多次抽样中样本均值间的离散程度，反映了样本均值对总体均值的代表性，用于推论统计。

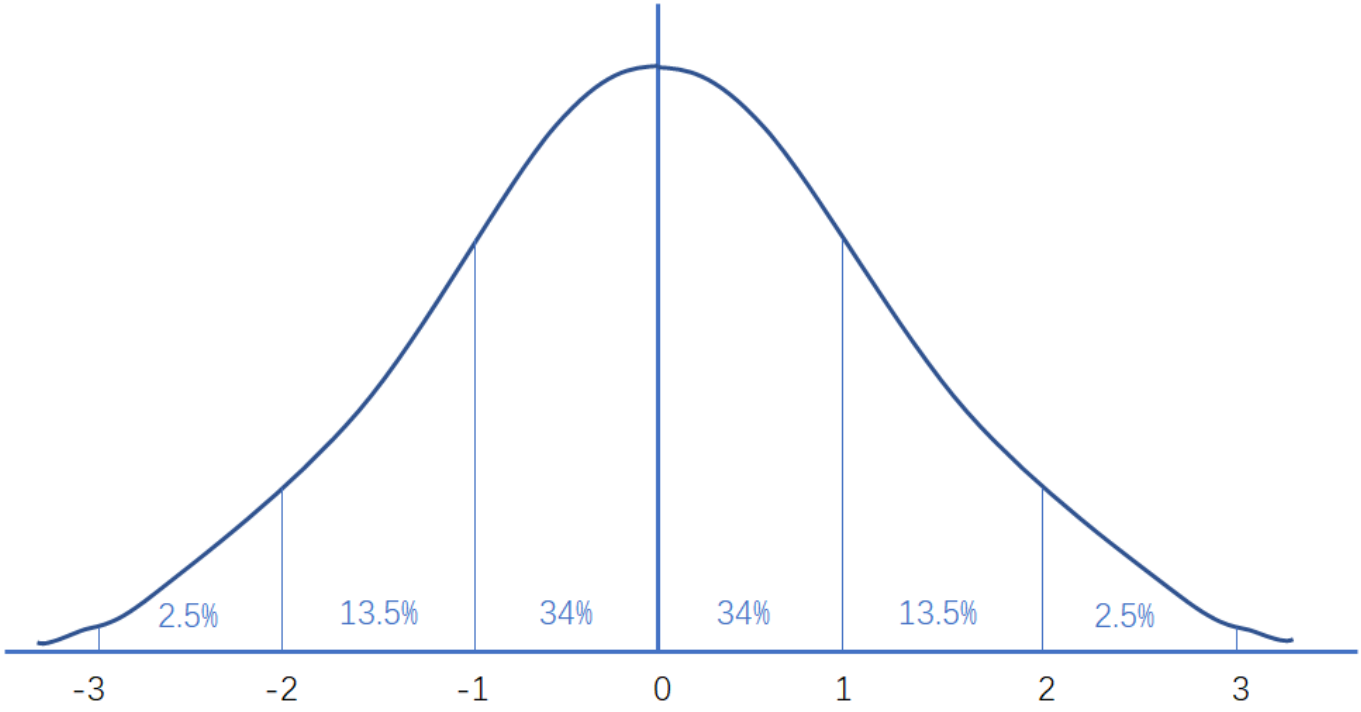
标准误差的具体使用

标准误差经常会被用于拿出一部分样品去判断整体产品线的产品质量，或者判断一个事情是不是属于常见范围。

比如说我们常见的六希格玛（Six Sigma），其实就是指所有的产品质量问题需要控制在 6 个标准误差里面。你听到的产品质量或者运维故障控制在 3 个 9 或者 5 个 9，说的也是误差范围。5 个 9 的意思就是 99.99966% 的产品是没有品质问题的。

这个是 99.99966% 怎么算出来的呢？这就涉及我们 06 讲里正态分布的知识，你要是记不清了，可以回过头再复习一下。

比如说我们用下面这个图做质量控制，那么这些值就是标准误差范围。例如，我们说在一个标准误差范围里，大概就是图里面的 68.3%；两个标准误差范围里也就是距离均值（标准件）的 95.4%；三个标准误差就是 99.7%；6 个标准误差（也就是 6-sigma）也就代表着要控制到在生产的产品中，有 99.99966% 的产品是没有品质问题的（每一百万件产品中只有 3.4 件有缺陷）。



所以说我们从标准误差来看，系统的稳定性要保证 5 个 9、6 个 9 或者说我们开发的代码的质量控制是 6-sigma，这个质量就非常好了。这么说可能你还没有感觉，我再给你做个比喻，帮助你理解一下。

帅哥美女其实在社会里面是非常难得一见的，毕竟我们绝大部分都是普通人。我们先假设人类美丽和帅的程度随机分布（整容的人没那么多），如果你每天见到一个美女（帅哥），那么有如下公式成立：

1 个标准误差的美女约为 3 天一遇；

2 个标准误差的美女为约 22 天一遇；

3 个标准误差的美女约为 370 天一遇；

4 个标准误差的美女约为 43 年一遇；

5 个标准误差的美女约为 4779 年一遇；

6 个标准误差的美女约为 139 万年一遇；

7 个标准误差的概率约为 10 亿年一遇。

这样看，你就知道 6 个标准误差有多么严格了。你下次遇到一个特别美丽的女孩子，你觉得她是万年一遇的美女的时候，你可以和她说：“啊，你是 6 个标准差一遇的美女

啊！”这肯定比说“你真美丽”要有深度得多，她肯定会对你学识非常敬仰（开个玩笑，真说估计会被打出去）。这样你应该知道标准误差是一个什么含义了。

小结

好，回顾一下今天的内容，今天主要给你讲了两个概念：标准差和标准误差。

标准差针对已经发生的事情，它是平均值的一个补充标准。而标准误差是多次抽样当中对样本离散程度的描述，用于推论中使用。在后面的内容里，我们还会用到这两兄弟来评估和衡量一个算法的稳定性以及实现结果的好坏。

看一个人、一个企业、一个投资产品靠不靠谱，除了人办事情的成功率、企业收入平均值和产品的盈利率，你还要看它标准差是怎样的。有可能这个所谓的“成功人士”只成功了一次，赚了一大笔钱，但是其实别的事他都失败了，那说明这人的标准差很大，有可能他就是靠运气，不太靠谱。我们中国人其实是比较喜欢“中庸”的这种感觉，用标准差的视角来看，就是自己做事做人的标准差要小。

对标准误差来说，我送你一个成语，叫做“严于律己，宽于待人”。前半句是指我们在工作和生活中，要尽量少出错，甚至是不犯错，这样不仅做事漂亮，领导喜欢，而且这种不断追求完美的理念，会一直推着我们往前跑。你可以试试，把六西格玛的思想不仅用在工作中，也用在生活里，对自己高标准、严要求一段时间，相信你会获得更进一步的成长。后半句是说，躺平无罪，奋斗有理。我们可以用六个标准误差来要求自己，但是别人也有用一个标准误差要求自己的自由。

如果用一句话来概括，**希望你尽量把自己做人做事的标准差变小，提高对自己的标准差预期。**

数据给你一双看透本质的眼睛，数据知识学无止境，让我们一起持续学习，一起共勉。

课后思考

你过去遇到过什么黑天鹅事件吗？从你的角度看，它是几个标准误差的范围呢？欢迎你在评论区分享你的想法，我们一起提高。

附录：方差及标准差公式

$$方差：Var(X) = E\big(X - E(X)\big)^2 = \begin{cases} \sum_i \big(x_i - E(X)\big)^2 p(x_i), & 离散情况 \\ \int_{-\infty}^{\infty} \big(x - E(X)\big)^2 p(x)dx, & 连续情况 \end{cases}$$

$$标准差：\sigma(X) = \sqrt{Var(X)}$$

分享给需要的人，Ta订阅后你可得 20 元现金奖励

赞 1 提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇 07 | 散点图和相关性：怎样快速从数据当中找到规律？

下一篇 09 | 数据抽样：大数据来了还需要抽样么？

精选留言 (3)

写留言

进化菌 神

2021-08-13

标准差让我想起了一个词，叫“长期主义”，大部分时间变化不大的情况下长期做一件事情。

1

2021-08-13

请问老师，我们开发的代码的质量控制是 6-sigma，这个要怎么计算或者推断呢？
代码开发完成之后，进行单元测试，接口测试，然后QA测试完成之后，发布到线上。线上发布三个月出了两次问题，算个比例么？

展开

作者回复: 好问题，这是一个管理问题，不同阶段考核不同，编程质量一般看单元测试和内部QA测试，整体产品质量看上线问题，因为上线后问题不仅仅是研发问题，产品设计，环境问题都算

Shark

2021-08-13

标准误差（Standard error）= 多次抽样中样本均值间的离散程度，反映了样本均值对总体均值的代表性，用于推论统计。 对样本的数量有要求吗？

展开

作者回复: 有的，这个可以根据正态分布或者数据分布来进行统计，你可以参考第6节课里推算样本数量的部分，我也加入到这个课程里参考里面。